Lecture 25:

Regression

Prediction

Suppose Y is some r.v.

What's the "best" guess about Y's value?

By "best" we mean minimizes mean squared error.

Theorem: the minimizer of $\mathbb{E}[(Y-\alpha)^2]$ is

$$\alpha = \mathbb{E}Y.$$

the minimum mean squared error is

Proof: $\mathbb{E}[(Y-\alpha)^2] = \mathbb{E}[Y^2] - 2\alpha\,\mathbb{E}[Y] + \alpha^2$

This is a parabola in α $\quad a\alpha^2 + b\alpha + c$

with minimal value at $\alpha = \frac{-b}{2a} = \frac{2\,\mathbb{E}Y}{2} = \mathbb{E}Y.$

Prediction

Suppose X, Y are some r.v.

Suppose we learned X's value.

What's the best guess for Y's value?

Recap

Multiple Random Variables

Joint Distribution: If X and Y are two r.v.s over the same probability space then their joint distribution is defined as

$$\{(a, b, \Pr[X=a, Y=b]) : a \in \mathrm{range}(X), b \in \mathrm{range}(Y)\}$$

Marginal Distributions

Marginal for X: $\Pr[X=a] = \sum_{b \in \mathrm{range}(Y)} \Pr[X=a, Y=b]$

Marginal for Y: $\Pr[Y=b] = \sum_{a \in \mathrm{range}(X)} \Pr[X=a, Y=b]$

Recap

Multiple Random Variables

Joint Distribution:

$$\{(a, b, \Pr[X=a, Y=b]) : a \in \text{range}(X), b \in \text{range}(Y)\}$$

Marginal for X: $\Pr[X=a] = \sum_{b \in \text{range}(Y)} \Pr[X=a, Y=b]$

Marginal for Y: $\Pr[Y=b] = \sum_{a \in \text{range}(X)} \Pr[X=a, Y=b]$

Example:

X \ Y	1	2	3
1	0	0.1	0.2
2	0.3	0	0
3	0.1	0.2	0.1

$\Pr[X=1] =$

$\Pr[Y=3] =$

$\Pr[X=1 \mid Y=3] =$

Conditional Expectation

Def'n: Let X and Y be r.v.s over Ω.

The conditional expectation of Y given X is defined as

$$\mathbb{E}[Y|X] = g(X)$$

where

$$g(x) = \mathbb{E}[Y|X=x] = \sum_{y} y \cdot \Pr[Y=y \mid X=x].$$

Note: $\mathbb{E}[Y|X]$ is a r.v. that is a fnc. of X.

$\forall x$: $\mathbb{E}[Y|X=x]$ is a number.

Properties of Conditional Expectation

$$\mathbb{E}[Y \mid X=x] = \sum_y y \cdot \Pr[Y=y \mid X=x]$$

1. If X, Y indep. $\Longrightarrow$ $\mathbb{E}[Y \mid X] = \mathbb{E}[Y]$.
2. $\mathbb{E}[aY+b \mid X] = a \cdot \mathbb{E}[Y \mid X] + b$
3. $\forall$ fnc $h(\cdot)$ $\quad \mathbb{E}[h(X)Y \mid X] = h(X) \cdot \mathbb{E}[Y \mid X]$.
4. $\mathbb{E}[\mathbb{E}[Y \mid X]] = \mathbb{E}[Y]$
5. $\forall$ fnc $h(\cdot)$ $\quad \mathbb{E}[h(X)\,\mathbb{E}[Y \mid X]] = \mathbb{E}[h(X) \cdot Y]$

Proof:

(1) true, by definition. (2) true, by linearity of expectation.

(3) For any x, $\quad \mathbb{E}[h(X) \cdot Y \mid X=x] = h(x) \cdot \mathbb{E}[Y \mid X=x]$.

(4) $$\mathbb{E}_X[\mathbb{E}_Y[Y \mid X]] = \sum_x \Pr[X=x] \cdot \sum_y y \cdot \Pr[Y=y \mid X=x]$$
$$= \sum_y y \cdot \underbrace{\sum_x \Pr[X=x] \cdot \Pr[Y=y \mid X=x]}_{\Pr[Y=y]} = \mathbb{E}[Y].$$

5. For any $h(\cdot)$ $\quad \underset{X}{\mathbb{E}}\left[h(X)\underset{Y}{\mathbb{E}}[Y \mid X]\right] = \underset{X,Y}{\mathbb{E}}[h(X)Y]$.

Proof:

$$\underset{X}{\mathbb{E}}\left[h(X)\cdot\underset{Y}{\mathbb{E}}[Y \mid X]\right] =$$

$$= \sum_{x} \Pr[X=x]\cdot h(x)\cdot \sum_{y} \Pr[Y=y \mid X=x]\cdot y \,.$$

$$= \sum_{x,y} h(x)\cdot y\cdot \Pr[X=x, Y=y]$$

$$= \underset{X,Y}{\mathbb{E}}[h(X)\cdot Y].$$

Corollary: For all $h(\cdot)$ $\quad \mathbb{E}[(Y-\mathbb{E}[Y|X])\cdot h(X)]=0$

Proof:

$$\mathbb{E}[(Y-\mathbb{E}(Y|X))h(X)] =$$

$$= \mathbb{E}[Y\cdot h(X)] - \mathbb{E}[\mathbb{E}[Y|X]\,h(X)]$$

$$= \mathbb{E}[Y\cdot h(X)] - \mathbb{E}[Y\cdot h(X)].$$

Theorem: Let X, Y be two r.v.s over Ω.

The best predictor of Y from X (minimizes mean squared error)

is $g(X) = \mathbb{E}[Y|X]$.

Proof: Let $h(X)$ be any function of X.

$$\mathbb{E}[(Y-h(X))^2] = \mathbb{E}[(Y-g(X)+g(X)-h(X))^2]$$

$$= \mathbb{E}[(Y-g(X))^2] + \underbrace{\cancel{2\mathbb{E}[(Y-g(X))(g(X)-h(X))]}}_{0} + \underbrace{\mathbb{E}[(g(X)-h(X))^2]}_{\geq 0}$$

$$\geq \mathbb{E}[(Y-g(X))^2]$$

Alternative Proof: For any $x \in \text{range}(X)$, given $X=x$

the best predictor of Y is $\mathbb{E}[Y|X=x]$ from earlier.

Linear Regression

So far, we've seen:

- If we want to guess Y without knowing anything else
  best guess is $\mathbb{E}Y$.
- If we make some observation X related to Y:
  best guess is $g(x) = \mathbb{E}[Y|X]$.

The latter is optimal but can be complicated.

What if we want a simpler fnc of X explaining Y.

For example: a linear function.

Motivation: Statistics

In real-life applications, we don't necessarily know the joint dist. of X, Y.

We can get estimates for $\mathbb{E}X$, $\mathbb{E}Y$, etc. from observations.

To estimate $\mathbb{E}[Y \mid X = x]$ we need samples such that $X = x$, but typically we'll have few or no such examples.

For a simpler model, $\alpha X + \beta$, we can use all the samples to get a good estimate of the two parameters: α, β.

Theorem: The best linear predictor of Y as a function of X is $\mathbb{E}[Y] + \frac{\operatorname{cov}(X,Y)}{\operatorname{var}(X)} \cdot (X - \mathbb{E}X)$.

Proof: We'll first consider predicting $\hat{Y} = Y - \mathbb{E}Y$ from $\hat{X} = X - \mathbb{E}X$.

$$\min_{\alpha,\beta} \mathbb{E}\left[\left(\hat{Y} - (\alpha\hat{X} + \beta)\right)^2\right] =$$

$$= \min_{\alpha,\beta} \mathbb{E}[\hat{Y}^2] - 2\mathbb{E}[\hat{Y}(\alpha\hat{X}+\beta)] + \mathbb{E}[(\alpha\hat{X}+\beta)^2]$$

$$= \min_{\alpha,\beta} \mathbb{E}[\hat{Y}^2] - 2\alpha\,\mathbb{E}[\hat{X}\hat{Y}] + \alpha^2\,\mathbb{E}[\hat{X}^2] + \beta^2$$

Solution: $\beta = 0$, $\alpha = \frac{\mathbb{E}[\hat{X}\cdot\hat{Y}]}{\mathbb{E}[\hat{X}^2]} = \frac{\operatorname{cov}(X,Y)}{\operatorname{Var}(X)}$

$\Rightarrow$ Best linear predictor for $Y - \mathbb{E}Y$ is $\alpha \cdot (X - \mathbb{E}X)$

$\Rightarrow$ " " " " Y is $\mathbb{E}[Y] + \alpha \cdot (X - \mathbb{E}X)$

Theorem: The best linear predictor of Y as a function of X is

$$\mathbb{E}Y + (X - \mathbb{E}X) \cdot \frac{\operatorname{Cov}(X,Y)}{\operatorname{Var}(X)}.$$

Corollary: The minimum $\mathbb{E}[(Y - \ell(X))^2] = \operatorname{Var}(Y) \cdot (1 - \operatorname{Corr}(X,Y)^2)$.

Proof: By the theorem $\ell(X) = \mathbb{E}Y + (X - \mathbb{E}X) \frac{\operatorname{Cov}(X,Y)}{\operatorname{Var}(X)}$

$$\mathbb{E}[(Y - \ell(X))^2] = \mathbb{E}\left[\left((Y - \mathbb{E}Y) - (X - \mathbb{E}X)\frac{\operatorname{Cov}(X,Y)}{\operatorname{Var}(X)}\right)^2\right]$$

$$= \mathbb{E}[(Y - \mathbb{E}Y)^2] - 2\frac{\operatorname{Cov}(X,Y)}{\operatorname{Var}(X)} \mathbb{E}[(Y - \mathbb{E}Y)(X - \mathbb{E}X)]$$

$$+ \left(\frac{\operatorname{Cov}(X,Y)}{\operatorname{Var}(X)}\right)^2 \cdot \mathbb{E}[(X - \mathbb{E}X)^2]$$

$$= \operatorname{Var}(Y) - \frac{\operatorname{Cov}(X,Y)^2}{\operatorname{Var}(X)}$$

$$= \operatorname{Var}(Y) - \operatorname{Var}(Y) \cdot \frac{\operatorname{Cov}(X,Y)^2}{\operatorname{Var}(X)\operatorname{Var}(Y)} = \operatorname{Var}(Y) \cdot (1 - \operatorname{Corr}(X,Y)^2).$$

Theorem: The best linear predictor of Y as a function of X is

$$\mathbb{E}Y + (X - \mathbb{E}X) \cdot \frac{\operatorname{cov}(X,Y)}{\operatorname{Var}(X)}.$$

Corollary: The minimum $\mathbb{E}\left[(Y - \ell(X))^2\right] = \operatorname{Var}(Y) \cdot (1 - \operatorname{corr}(X,Y)^2)$.

This is what we mean by

"X explains 80% of the variance of Y"

$$\Updownarrow$$

$$\operatorname{corr}(X,Y)^2 = 0.8.$$

Minimizing Given Data

Suppose you get samples $(x_1, y_1), \dots, (x_n, y_n)$ from the joint distribution, and you want to minimize

$$\sum_{i=1}^{n} \left(y_i - \alpha x_i + \beta\right)^2$$

You'll get: $\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i$

$$\alpha = \frac{\frac{1}{n} \sum_{i=1}^{n} x_i y_i - \bar{x} \cdot \bar{y}}{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

(estimate to $\frac{Cov(X,Y)}{Var(X)}$)

$$\beta = \bar{y} - \alpha \cdot \bar{x}$$

(estimate to $\mathbb{E}Y - \alpha \mathbb{E}X$)